# Convergence rates to deflation of simple shift strategies

Ricardo S. Leite, Nicolau C. Saldanha and Carlos Tomei

August 16, 2010

**Abstract**

The computation of eigenvalues of real symmetric tridiagonal matrices frequently proceeds by a sequence of $QR$ steps with shifts. We introduce *simple shift strategies*, functions $\sigma$ satisfying natural conditions, taking each $n \times n$ matrix $T$ to a real number $\sigma(T)$. The strategy specifies the shift to be applied by the $QR$ step at $T$. Rayleigh and Wilkinson's are examples of simple shift strategies. We show that if $\sigma$ is continuous then there exist initial conditions for which deflation does not occur, i.e., subdiagonal entries do not tend to zero. In case of deflation, we consider the rate of convergence to zero of the $(n, n-1)$ entry: for simple shift strategies this is always at least quadratic. If the function $\sigma$ is smooth in a suitable region and the spectrum of $T$ does not include three consecutive eigenvalues in arithmetic progression then convergence is cubic. This implies cubic convergence to deflation of Wilkinson's shift for generic spectra. The study of the algorithm near deflation uses *tubular coordinates*, under which $QR$ steps with shifts are given by a simple formula.

**Keywords:** Isospectral manifold, Deflation, Wilkinson's shift, Shifted $QR$ algorithm.

**MSC2010-class:** 65F15; 37N30.

## 1 Introduction

Let $\mathcal{T}$ be the vector space of real symmetric tridiagonal matrices. Among the standard algorithms to compute eigenvalues of matrices in $\mathcal{T}$ are $QR$ steps with different shift strategies: Rayleigh and Wilkinson are familiar examples (excellent references are [15], [4], [13]). In this paper, we consider a more general context: we define simple shift strategies, which include the examples above and more, and discuss subtle aspects of their asymptotic behavior.

More precisely, given a matrix $T \in \mathcal{T}$ and $s \in \mathbb{R}$, write $T - sI = QR$, if possible, for an orthogonal matrix $Q$ and an upper triangular matrix $R$ with positive diagonal entries. A *shifted $QR$ step* is $\Phi(T, s) = Q^*TQ$. As is well known, shifted $QR$ steps preserve spectrum and shape. For a real matrix with simple spectrum $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, let $\mathcal{T}_\Lambda \subset \mathcal{T}$ be the set of matrices similar to $\Lambda$. A *simple shift strategy* is a function $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$ satisfying the following two properties.

(I) For all $T \in \mathcal{T}_\Lambda$, $\sigma(E_nTE_n) = \sigma(T)$, where $E_n = \operatorname{diag}(1, 1, \dots, 1, -1)$.

(II) There exists $C_\sigma > 0$ such that for all $T \in \mathcal{T}_\Lambda$ there is an eigenvalue $\lambda_i$ with $|\sigma(T) - \lambda_i| \le C_\sigma |b(T)|$, where $b(T) = (T)_{(n,n-1)}$.

This definition excludes algorithms which employ multi-shifts and extrapolation techniques, but accomodates the usual Rayleigh and Wilkinson's strategies.

For technical reasons, we prefer the *signed* variant $\Phi_\star(T,s) = Q_\star^* T Q_\star$, where now $T - sI = Q_\star R_\star$, the orthogonal matrix $Q_\star$ has positive determinant and only the first $n-1$ diagonal entries of the upper triangular matrix $R_\star$ are required to be positive. It is easy to see that either $\Phi(T,s) = \Phi_\star(T,s)$ or $\Phi(T,s) = E_n \Phi_\star(T,s) E_n$ (notice that $E_n T E_n$ is obtained from $T \in \mathcal{T}$ by changing the signs of the entries $(n, n-1)$ and $(n-1, n)$). As we shall see, the signed step is smoothly defined on a larger domain, and convergence issues for both kinds of step iterations are essentially equivalent.

Simple shift strategies prescribe shifts: set $F_s(T) = \Phi_\star(T,s)$ and $F_\sigma(T) = F_{\sigma(T)}(T)$. Algorithms iterate $F_\sigma$ aiming at deflation, i.e., obtaining a matrix $T \in \mathcal{T}_\Lambda$ with small $|b(T)|$. The *deflation set* (resp. *neighborhood*) is the set $\mathcal{D}_{\Lambda,0} \subset \mathcal{T}_\Lambda$ (resp. $\mathcal{D}_{\Lambda,\epsilon} \subset \mathcal{T}_\Lambda$) consisting of matrices $T$ for which $b(T) = 0$ (resp. $|b(T)| \le \epsilon$). A simple shift strategy $\sigma$ is *deflationary* if for any $T \in \mathcal{T}_\Lambda$ and any $\epsilon > 0$ there exists $k$ for which $F_\sigma^k(T) \in \mathcal{D}_{\Lambda,\epsilon}$. As is well known, Rayleigh's strategy is not deflationary and Wilkinson's is: our first result provides a context for these facts.

**Theorem 1** *A continuous shift strategy $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$ is not deflationary.*

We are thus led to consider the *singular support* $\mathcal{S}_\sigma \subset \mathcal{T}_\Lambda$ of a shift strategy $\sigma$, i.e., the minimal closed subset of $\mathcal{T}_\Lambda$ on whose complement $\sigma$ is smooth. For Rayleigh $\mathcal{S}_\sigma$ is empty; for Wilkinson's strategy it consists of the matrices $T \in \mathcal{T}_\Lambda$ with $(T)_{n-1,n-1} = (T)_{n,n}$.

Numerical evidence brings up the question of whether the rate of convergence to zero of the sequence $b(F_\sigma^k(T))$ is cubic, in the sense that there is a constant $C$ such that $|b(F_\sigma^{k+1}(T))| \le C |b(F_\sigma^k(T))|^3$ for large $k$. It turns out that, for any shift strategy $\sigma$, in an appropriate neighborhood of the deflation set $\mathcal{D}_{\Lambda,0}$, each iteration of $F_\sigma$ squeezes the $(n, n-1)$ entry quadratically. Away from the singular support $\mathcal{S}_\sigma$, squeezing is cubic.

**Theorem 2** *For $\epsilon > 0$ small enough, the deflation neighborhood $\mathcal{D}_{\Lambda,\epsilon}$ is invariant under $F_\sigma$. There exists $C > 0$ such that, for all $T \in \mathcal{D}_{\Lambda,\epsilon}$, $|b(F_\sigma(T))| \le C |b(T)|^2$. Also, given a compact set $\mathcal{K} \subset \mathcal{D}_{\Lambda,\epsilon}$ disjoint from $\mathcal{S}_\sigma \cap \mathcal{D}_{\Lambda,0}$, there exists $C_\mathcal{K} > 0$ such that, for all $T \in \mathcal{K}$, $|b(F_\sigma(T))| \le C_\mathcal{K} |b(T)|^3$.*

Cubic convergence does not hold in general for Wilkinson's strategy. In [8], for $\Lambda = \mathrm{diag}(-1, 0, 1)$, we construct a Cantor-like set $\mathcal{X} \subset \mathcal{T}_\Lambda$ of unreduced initial conditions for which the rate of convergence is strictly quadratic. Sequences starting at $\mathcal{X}$ converge to a reduced matrix which is not diagonal. For Rayleigh's shift, on the other hand, convergence is always cubic within invariant deflation neighborhoods.

A matrix $T \in \mathcal{T}$ with simple spectrum is *a.p. free* if it does not have three eigenvalues in arithmetic progression and *a.p.* otherwise. For a.p. free spectra, the situation is very nice: cubic convergence is essentially uniform on $\mathcal{T}_\Lambda$.

**Theorem 3** *Let $\Lambda$ be an a.p. free matrix and $\sigma$ a shift strategy for which diagonal matrices do not belong to $\mathcal{S}_\sigma$. Then there exist $\epsilon > 0$, $C > 0$ and $K > 0$ such that the deflation neighborhood $\mathcal{D}_{\Lambda,\epsilon}$ is invariant under $F_\sigma$. Also, for any $T \in \mathcal{D}_{\Lambda,\epsilon}$, the sequence $(F_\sigma^k(T))$ converges to a diagonal matrix and the set of positive integers $k$ for which $|b(F_\sigma^{k+1}(T))| > C |b(F_\sigma^k(T))|^3$ has at most $K$ elements.*

Still, the finite set of points in which the cubic estimate does not hold may occur arbitrarily late along the sequence $(F_\sigma^k(T))$.

An a.p. matrix is *strong a.p.* if it contains three consecutive eigenvalues in arithmetic progression and *weak a.p.* otherwise. Under very mild additional hypothesis, $b(T)$ converges to zero at a cubic rate also for weak a.p. matrices. Let $\mathcal{C}_{\Lambda,0} \subset \mathcal{T}_\Lambda$ be the set of matrices $T$ for which $(T)_{n,n-1} = (T)_{n-1,n-2} = 0$.

**Theorem 4** *Let $\Lambda$ be a weak a.p. matrix and $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$ a shift strategy for which $\mathcal{C}_{\Lambda,0}$ and $\mathcal{S}_\sigma$ are disjoint. Then there exists $\epsilon > 0$ such that the deflation neighborhood $\mathcal{D}_{\Lambda,\epsilon}$ is invariant under $F_\sigma$ and, for all unreduced $T \in \mathcal{D}_{\Lambda,\epsilon}$, the sequence $(b(F_\sigma^k(T)))$ converges to zero at a rate which is at least cubic. More precisely, for each unreduced $T \in \mathcal{D}_{\Lambda,\epsilon}$ there exist $C_T, K_T > 0$ such that, for all $k > K_T$, we have $|b(F_\sigma^{k+1}(T))| \leq C_T |b(F_\sigma^k(T))|^3$.*

In particular, the convergence of Wilkinson's strategy is cubic for weak a.p. matrices. However, uniformity in the sense of Theorem 3 is not guaranteed and the constants $C_T$ and $K_T$ depend on $T$. As in the case of the spectrum $\{-1, 0, 1\}$, we conjecture that if $\Lambda$ is strong a.p. then there exists $\mathcal{X} \subset \mathcal{T}_\Lambda$ of Hausdorff codimension 1 of initial conditions $T$ for which the rate of convergence is strictly quadratic.

The proofs of the above results depend on few basic ideas. Signed shifted steps $F_s(T)$ are shown to be well defined for unreduced matrices and in an open neighborhood of the deflation set $\mathcal{D}_{\Lambda,0}$. The compact set (manifold) $\mathcal{D}_{\Lambda,0}$ splits into connected components $\mathcal{D}^i_{\Lambda,0}$ consisting of matrices $T \in \mathcal{T}_\Lambda$ with $(T)_{n,n} = \lambda_i$. For small $\epsilon$, the deflation neighborhood splits into components $\mathcal{D}^i_{\Lambda,\epsilon}$ which are thickenings of $\mathcal{D}^i_{\Lambda,0}$.

*Tubular coordinates* provide a good understanding both of tubular neighborhoods $\mathcal{D}^i_{\Lambda,\epsilon}$ and of shifted $QR$ steps within these sets. A previous unpublished version of this paper ([9]) uses instead *bidiagonal coordinates*, defined in [7], to prove some of the results presented here for Wilkinson's shift; these coordinates are also used in [7] to prove the cubic convergence of Rayleigh's shift. Bidiagonal coordinates consist of very explicit charts on the manifold $\mathcal{T}_\Lambda$. On both coordinate systems, shifted $QR$ steps are very simple.

Sufficiently thin deflation neighborhoods are invariant under steps $F_\sigma$. Theorem 1 then becomes a connectivity argument: in a nutshell, there must be separatrices in order to permit deflation to different components $\mathcal{D}^i_{\Lambda,0}$.

Steps $F_\sigma$ are smooth whenever the shift strategy is, i.e., for $T \in \mathcal{D}^i_{\Lambda,\epsilon} \smallsetminus \mathcal{S}_\sigma$. At matrices $T_0 \in \mathcal{D}^i_{\Lambda,0}$ on which $F_\sigma$ is smooth, the map $T \mapsto b(F_\sigma(T))$ has zero gradient. The symmetry of the shift strategy (condition (I)) yields a cubic Taylor expansion and therefore an estimate $|b(F_\sigma(T))| \leq C|b(T)|^3$, settling Theorem 2.

*Height functions* $H : \mathcal{D}^i_{\Lambda,\epsilon} \to \mathbb{R}$ are used for further study of the sequence $(F_\sigma^k(T))$. More precisely, for steps $s$ near $\lambda_i$, $H_i(F_s(T)) > H_i(T)$ provided $T \in \mathcal{D}^i_{\Lambda,\epsilon}$ is not diagonal. A compactness argument then bounds the number of iterations for which $F_\sigma^k(T)$ stays close to the singular support $\mathcal{S}_\sigma$: this is essentially Theorem 3.

For a.p. spectra, the situation is subtler. As numerical analysts know, shift strategies usually define sequences of matrices which, asymptotically, not only isolate an eigenvalue at the $(n, n)$ position but also isolate, at a slower rate, a second eigenvalue at the $(n-1, n-1)$ position. This does not happen for the example in [8] where $(F_\sigma^k(T))_{n,n}$ tends to the center of a three-term arithmetic progression of eigenvalues and $(F_\sigma^k(T))_{n-1,n-2}$ stays bounded away from 0. On the other hand, Theorem 4 tells us that the weak a.p. hypothesis together with an appropriate smoothness condition guarantee cubic convergence.

In Section 2 we list the basic properties of the signed shifted $QR$ step. Simple shift strategies are introduced in Section 3, and the standard examples are shown to satisfy the definition. We define the deflation set $\mathcal{D}_{\Lambda,0}$ and neighborhood $\mathcal{D}_{\Lambda,\epsilon}$ in Section 4 and then set up tubular coordinates. The local theory of steps $F_s$ near $\mathcal{D}_{\Lambda,0}$ and the proof Theorem 2 are presented in Section 5. Section 6 is dedicated to Theorem 1. In Section 7 we construct the height functions $H$ and then prove Theorem 3. The convergence properties of a.p. matrices in Theorem 4 are proved in Section 8. We finally present in Section 9 two counterexamples to natural but incorrect strengthenings of Theorems 3 and 4.

The authors are very grateful for the abundant contributions of several readers of this work and its previous versions. The authors acknowledge support from CNPq, CAPES, IM-AGIMB and FAPERJ.

## 2 $QR$ iteration with shift and a variation

For a matrix $M$, the $QR$ factorization is $M = QR$ for an orthogonal matrix $Q$ and an upper triangular matrix $R$ with positive diagonal. As usual, let $SO(n)$ denote the set of orthogonal matrices with determinant equal to 1. The $Q_\star R_\star$ *factorization*, instead, is $M = Q_\star R_\star$, for $Q_\star \in SO(n)$ and $R_\star$ an upper triangular matrix with $(R_\star)_{i,i} > 0$, $i = 1, \ldots, n-1$. A real $n \times n$ matrix $M$ is *almost invertible* if its first $n-1$ columns are linearly independent: notice that almost invertible matrices are dense within $n \times n$ matrices and form an open set. The diagonal matrix $E_n$ is such that $(E_n)_{i,i}$ is 1 for $i < n$ and $-1$ for $i = n$.

**Proposition 2.1** *An almost invertible real matrix $M$ admits a unique $Q_\star R_\star$ factorization, with $Q_\star$ and $R_\star$ depending smoothly on $M$. If $M$ is invertible, it admits unique (smooth) factorizations $M = QR = Q_\star R_\star$. If $\det M > 0$, the factorizations are equal, i.e., $Q = Q_\star$ and $R = R_\star$. If $\det M < 0$, $Q = Q_\star E_n$ and $R = E_n R_\star$. If $\det M = 0$, $(R_\star)_{n,n} = 0$.*

**Proof:** Let $M$ be almost invertible. Applying Gram-Schmidt with positive normalizations on its first $n-1$ columns we obtain the first $n-1$ columns of both $Q$ and $R$, as well as those of $Q_\star$ and $R_\star$. The last column $v = Q_\star e_n$ of $Q_\star$ is already well defined, by orthonormality and the fact that $\det Q_\star = 1$. Now, set $R_\star = M(Q_\star)^*$. The positivity of $R_{n,n}$ specifies whether the last column of $Q$ is $v$ or $-v$. Smoothness is clear by construction.

If $M$ is invertible, $\det M = \det Q_\star \det R_\star$ implies that the last diagonal entry of $R_\star$ has the same sign of $\det M$: the relations between the factorizations then follow. If $M$ is not invertible, the relation among determinants implies $(R_\star)_{n,n} = 0$. ■

Let $\mathcal{T}$ denote the real vector space of $n \times n$ real, symmetric, tridiagonal matrices endowed with the norm $\|T\|^2 = \mathrm{tr}(T^2)$. For $T \in \mathcal{T}$, the *subdiagonal entries* of $T$ are $(T)_{i+1,i}$ for $i = 1, \ldots, n-1$. The lowest subdiagonal entry of $T$ is $b(T) = (T)_{n,n-1}$. If all subdiagonal entries of $T$ are nonzero, $T$ is an *unreduced* matrix; otherwise, $T$ is *reduced*. Notice that an unreduced tridiagonal matrix is almost invertible: indeed, the block formed by rows $2, \ldots, n$ and columns $1, \ldots, n-1$ is a an upper triangular matrix with nonzero diagonal entries, and therefore, invertible.

We consider the *shifted $QR$ step* and its *signed* counterpart,

$$\Phi(T, s) = Q^* T Q, \quad \Phi_\star(T, s) = Q_\star^* T Q_\star,$$

where $T - sI = QR$ and $T - sI = Q_\star R_\star$. The pair $(T, s) \in \mathcal{T} \times \mathbb{R}$ belongs to the natural (open, dense) domains $\mathrm{Dom}(\Phi)$ and $\mathrm{Dom}(\Phi_\star)$ if $T - sI$ is invertible and almost invertible, respectively: clearly, the functions $\Phi$ and $\Phi_\star$ are smooth in their domains.

**Lemma 2.2** *For $(T, s) \in \mathrm{Dom}(\Phi)$ (resp. $\mathrm{Dom}(\Phi_\star)$), we have $\Phi(T, s) \in \mathcal{T}$ (resp. $\Phi_\star(T, s) \in \mathcal{T}$). The spectra of $T$, $\Phi(T, s)$ and $\Phi_\star(T, s)$ are equal. In the appropriate domains, for $T - sI = QR = Q_\star R_\star$ and $i = 1, 2, \ldots, n-1$,*

$$(\Phi(T, s))_{i+1,i} = \frac{(R)_{i+1,i+1}}{(R)_{i,i}} (T)_{i+1,i}, \quad (\Phi_\star(T, s))_{i+1,i} = \frac{(R_\star)_{i+1,i+1}}{(R_\star)_{i,i}} (T)_{i+1,i}.$$

*Thus, the top $n-2$ subdiagonal entries of $T$, $\Phi(T, s)$ and $\Phi_\star(T, s)$ have the same sign; also, $\mathrm{sign}\,(T)_{n,n-1} = \mathrm{sign}\,(\Phi(T, s))_{n,n-1}$.*

**Proof:** We prove the statements for $\Phi_\star$; the others are then easy.

For a pair $(T,s) \in \mathrm{Dom}(\Phi) \subset \mathrm{Dom}(\Phi_\star)$, there are two expressions for $\Phi_\star(T,s)$:

$$\Phi_\star(T,s) = Q_\star^* T Q_\star = R_\star T R_\star^{-1}, \quad \text{where} \quad T - sI = Q_\star R_\star.$$

From the first equality, $\Phi_\star(T,s)$ is symmetric and from the second, $\Phi_\star(T,s)$ is an upper Hessenberg matrix so that $\Phi_\star(T,s) \in \mathcal{T}$ is similar to $T$. More generally, for $(T,s) \in \mathrm{Dom}(\Phi_\star)$ we still have

$$\Phi_\star(T,s) = Q_\star^* T Q_\star, \quad \Phi_\star(T,s) R_\star = R_\star T$$

and therefore $\Phi_\star(T,s) \in \mathcal{T}$ is similar to $T$. Compute the $(i+1,i)$ entry of the second equation above to obtain $(\Phi_\star(T,s))_{i+1,i}\ (R_\star)_{i,i} = (R_\star)_{i+1,i+1}\ (T)_{i+1,i}$, completing the proof. ■

The following result describes the behavior of $\Phi_\star$ at points not in $\mathrm{Dom}(\Phi)$, which will play an important role throughout the paper.

**Lemma 2.3** *If* $(T,s) \in \mathrm{Dom}(\Phi_\star) \setminus \mathrm{Dom}(\Phi)$ *then*

$$b(\Phi_\star(T,s)) = (\Phi_\star(T,s))_{n,n-1} = 0, \quad (\Phi_\star(T,s))_{n,n} = s.$$

*At a point* $(T,s) \in \mathrm{Dom}(\Phi_\star)$ *with* $b(T) = 0$ *and* $s = (T)_{n,n}$ *we have* $\mathrm{grad}(b \circ \Phi_\star) = 0$.

**Proof:** Since $T - sI = Q_\star R_\star = R_\star^* Q_\star^*$ is not invertible then $(R_\star)_{n,n} = 0$ and therefore $R_\star^* e_n = 0$. Thus $v = (Q_\star^*)^{-1} e_n = Q e_n$ satisfies $(T - sI)v = 0$. We then have $\Phi_\star(T,s) e_n = Q^* T Q e_n = Q^* T v = Q^*(sv) = s e_n$, proving the first claim. For the second claim, since $T - sI$ is almost invertible, $(R_\star)_{i,i} > 0$ for $i < n$. From the previous lemma,

$$(b \circ \Phi_\star)(T,s) = \frac{(R_\star)_{n,n}}{(R_\star)_{n-1,n-1}}\ b(T)\,;$$

if $b(T) = 0$ and $s = (T)_{n,n}$ then $(R_\star)_{n,n} = 0$ and $b \circ \Phi_\star$ is a product of two smooth functions, both zero, yielding $\mathrm{grad}(b \circ \Phi_\star) = 0$. ■

Recall that symmetrically changing the signs of subdiagonal entries does not change the spectrum of a matrix in $\mathcal{T}$. Let $\mathcal{E}$ denote the set of *signed diagonal matrices* with $\pm 1$ along the diagonal entries; in particular, $E_n \in \mathcal{E}$. The operation of changing subdiagonal signs, i.e., of conjugation by some $E \in \mathcal{E}$, behaves well with respect to $\Phi$ and $\Phi_\star$.

**Lemma 2.4** *Let* $E \in \mathcal{E}$. *The domains* $\mathrm{Dom}(\Phi)$ *and* $\mathrm{Dom}(\Phi_\star)$ *are invariant under conjugation by* $E$ *and*

$$\Phi(ETE, s) = E\Phi(T,s)E, \quad \Phi_\star(ETE, s) = E\Phi_\star(T,s)E.$$

*If* $\det(T - sI) > 0$ *then* $\Phi(T,s) = \Phi_\star(T,s)$; *if* $\det(T - sI) < 0$, $\Phi(T,s) = E_n \Phi_\star(T,s) E_n$; *if* $\det(T - sI) = 0$ *and* $(T,s) \in \mathrm{Dom}(\Phi_\star)$, *then* $b(\Phi_\star(T,s)) = 0$.

**Proof:** For $(T,s) \in \mathrm{Dom}(\Phi)$, the matrices $T - sI$ and $E(T - sI)E$ are both invertible. The $QR$ factorization $T - sI = QR$ yields $ETE - E(sI)E = (EQE)(ERE)$, preserving the positivity of the diagonal entries of the triangular part, so

$$\Phi(ETE, s) = (EQE)^* ETE (EQE) = EQ^* T Q E = E\Phi(T,s)E.$$

The argument is similar for $\Phi_\star$. The claims for $T - sI$ invertible follow from the relation between $Q$ and $Q_\star$ in Proposition 2.1; the case $\det(T - sI) = 0$ is a repetition of Lemma 2.3. ■

We are only interested in the case when the spectrum of $T$ is simple, since a double eigenvalue implies reducibility. Let $\Lambda$ be a real diagonal matrix with simple eigenvalues $\lambda_1 < \cdots < \lambda_n$. Define the *isospectral manifold*

$$\mathcal{T}_\Lambda = \{Q^*\Lambda Q, Q \in SO(n)\} \cap \mathcal{T},$$

the set of matrices in $\mathcal{T}$ similar to $\Lambda$. The set $\mathcal{T}_\Lambda \subset \mathcal{T}$ is a real smooth manifold ([14]; [7] describes an explicit atlas of $\mathcal{T}_\Lambda$). Since either version of shifted $QR$ step preserves spectrum, restriction defines smooth maps $\Phi : (\mathcal{T}_\Lambda \times \mathbb{R}) \cap \mathrm{Dom}(\Phi) \to \mathcal{T}_\Lambda$ and $\Phi_\star : (\mathcal{T}_\Lambda \times \mathbb{R}) \cap \mathrm{Dom}(\Phi_\star) \to \mathcal{T}_\Lambda$.

Still in $\mathcal{T}_\Lambda$, it is convenient to consider the *step* $F_s(T) = \Phi_\star(T, s)$. For $s$ not an eigenvalue of $\Lambda$, the domain of $F_s$ is $\mathcal{T}_\Lambda$. The natural domain for $F_{\lambda_i}$ instead is the *deflation domain* $\mathcal{D}^i_\Lambda$, the open dense subset of $\mathcal{T}_\Lambda$ of matrices $T$ for which $T - \lambda_i I$ is almost invertible. In other words, $T \in \mathcal{D}^i_\Lambda$ if and only if $\lambda_i$ is an eigenvalue of the lowest irreducible block of $T$.

The definition of the step $F_s$ differs from the usual one in that we use $\Phi_\star$ instead of $\Phi$. Given Lemma 2.4, considerations about deflation are unaffected and our choice has the advantage of being smooth (and well defined) in $\mathcal{D}^i_\Lambda$.

The ($i$-th) *deflation set* is

$$\mathcal{D}^i_{\Lambda,0} = \{T \in \mathcal{T}_\Lambda \mid b(T) = 0, (T)_{n,n} = \lambda_i\}.$$

Since the spectrum of $\Lambda$ is simple, $\mathcal{D}^i_{\Lambda,0} \subset \mathcal{D}^i_\Lambda$. Also, if $i \neq j$ then $\mathcal{D}^i_\Lambda \cap \mathcal{D}^j_{\Lambda,0} = \emptyset$. We saw in Lemma 2.3 that when the shift is taken to be an eigenvalue, a single step deflates a matrix, i.e., that the image of $F_{\lambda_i}$ is contained in $\mathcal{D}^i_{\Lambda,0}$: we shall see in Proposition 4.1 that this image is in fact equal to $\mathcal{D}^i_{\Lambda,0}$.

# 3 Simple shift strategies

Quoting Parlett [13], there are shifts for all seasons. The point of using a shift strategy is to accelerate deflation, ideally by choosing $s$ near an eigenvalue of $T$. A *simple shift strategy* is a function $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$ such that:

(I) for all $T \in \mathcal{T}_\Lambda$, $\sigma(E_n T E_n) = \sigma(T)$;

(II) there exists $C_\sigma > 0$ such that for all $T \in \mathcal{T}_\Lambda$ there is an eigenvalue $\lambda_i$ with $|\sigma(T) - \lambda_i| \leq C_\sigma |b(T)|$.

In particular, if $T \in \mathcal{D}^i_{\Lambda,0}$ then $\sigma(T) = \lambda_i$. The *step* associated with a (simple) shift strategy $\sigma$ is $F_\sigma$, defined by $F_\sigma(T) = F_{\sigma(T)}(T)$. The natural domain for $F_\sigma$ is the set of matrices $T$ for which $T - \sigma(T)I$ is almost invertible. From Section 2, it includes all unreduced matrices and open neighborhoods of each deflation set $\mathcal{D}^i_{\Lambda,0}$. We shall also see in Section 6 that it contains a dense open subset $\mathcal{U}_{\Lambda,\epsilon}$ of $\mathcal{T}_\Lambda$ invariant under $F_\sigma$. A more careful description of this domain will not be needed.

We leave to the reader the verification that *Rayleigh's shift* $\rho(T) = (T)_{n,n}$ is a simple shift strategy. Denote the bottom $2 \times 2$ diagonal principal minor of a matrix $T \in \mathcal{T}$ by $\hat{T}$: *Wilkinson's shift* $\omega(T)$ is the eigenvalue of $\hat{T}$ closer to $(T)_{n,n}$ (in case of draw, take the smallest eigenvalue).

**Lemma 3.1** *The function $\omega$ is a simple shift strategy with $C_\omega = 2\sqrt{2}$.*

**Proof:** Condition (I) follows from the fact that changing signs of off-diagonal entries of a $2\times2$ matrix does not change its spectrum. For (II), apply the Wielandt-Hoffman theorem to the $2 \times 2$ trailing principal minors of $T$ and $S = T - b(T)B$ to deduce that $|(T)_{n,n} - \omega(T)| \leq \sqrt{2}\, b(T)$. Again from Wielandt-Hoffman, now on $S$ as in Proposition 4.4, $|(T)_{n,n} - \lambda_i| \leq \sqrt{2}\, b(T)$ and $|\omega(T) - \lambda_i| \leq 2\sqrt{2}\, b(T)$. ■

Another example of shift strategy, the *mixed Wilkinson-Rayleigh strategy*, uses Wilkinson's shift unless the matrix is already near deflation, in which case we use Rayleigh's:

$$\sigma(T) = \begin{cases} \rho(T), & |(T)_{n,n-1}| < \epsilon, \\ \omega(T), & |(T)_{n,n-1}| \geq \epsilon; \end{cases}$$

here $\epsilon > 0$ is a small constant.

Shift strategies are not required to be continuous and $\omega$ is definitely not. For a shift strategy $\sigma$, let $\mathcal{S}_\sigma \subset \mathcal{T}_\Lambda$ be the *singular support* of $\sigma$, i.e., a minimal closed set on whose complement $\sigma$ is smooth. For example, $\mathcal{S}_\omega$ is the set of matrices $T \in \mathcal{T}_\Lambda$ for which the two eigenvalues $\omega_-(T)$ and $\omega_+(T)$ of $\hat{T}$ are equidistant from $(T)_{n,n}$, or, equivalently, for which $(T)_{n,n} = (T)_{n-1,n-1}$. The set $\mathcal{S}_\sigma$ will play an important role later.

We consider the phase portrait of $F_\omega$ for $3 \times 3$ matrices. In this case, the reader may check that the domain of $F_\omega$ is the full set $\mathcal{T}_\Lambda$. Let $\mathcal{J}_\Lambda \subset \mathcal{T}_\Lambda$ be set of *Jacobi matrices* similar to $\Lambda$, i.e., matrices $T \in \mathcal{T}_\Lambda$ with strictly positive subdiagonal entries. It is known ([2], [11]) that the closure $\bar{\mathcal{J}}_\Lambda \subset \mathcal{T}_\Lambda$ is diffeomorphic to a hexagon. The set $\bar{\mathcal{J}}_\Lambda$ is not invariant under $F_\omega$ but we may define $\tilde{F}_\omega(T)$ with $\tilde{F}_\omega : \bar{\mathcal{J}}_\Lambda \to \bar{\mathcal{J}}_\Lambda$ by dropping signs of subdiagonal entries of $F_\omega(T)$. As discussed above, this rather standard procedure is mostly harmless.

Two examples of $\tilde{F}_\omega$ are given in Figure 1, which represent $\bar{\mathcal{J}}_\Lambda$ for the $\Lambda = \mathrm{diag}(1, 2, 4)$ on the left and $\Lambda = \mathrm{diag}(-1, 0, 1)$ on the right. The vertices are the six diagonal matrices similar to $\Lambda$ and the edges consist of reduced matrices. Labels indicate the diagonal entries of the corresponding matrices. Three edges form $\mathcal{D}^i_{\Lambda,0} \cap \bar{\mathcal{J}}_\Lambda$: they alternate, starting from the bottom horizontal edge on both hexagons. The set $\mathcal{S} \cap \bar{\mathcal{J}}_\Lambda$ is indicated in both cases.

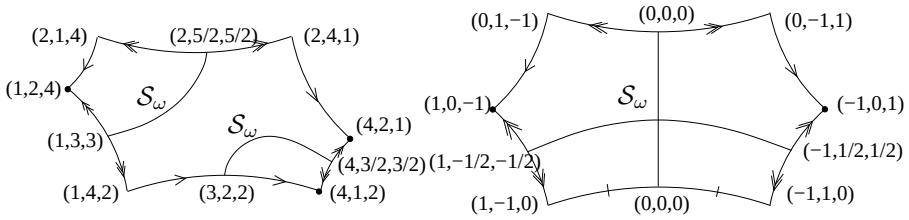


Figure 1: The phase space of Wilkinson's step for $n = 3$.

Vertices are fixed points of $\tilde{F}_\omega$ and boundary edges are invariant sets. A simple arrow indicates the motion of the points $\tilde{F}^k_\omega(T)$ along the edge. Points $T$ on an arc with a double arrow are taken to a diagonal matrix in a single step: the arc points to $\tilde{F}_\omega(T)$. Arcs marked with a transversal segment consist of fixed points of $\tilde{F}_\omega$.

Points on both sides of $\mathcal{S}_\omega$ are taken far apart: there is essentially a jump discontinuity along $\mathcal{S}_\omega$. From Theorem 2, the decay of the bottom subdiagonal entry under Wilkinson's step away from $\mathcal{S}_\omega \cap \mathcal{D}_{\Lambda,0}$ is cubic. As discussed in [8], near $\mathcal{S}_\omega \cap \mathcal{D}_{\Lambda,0}$ this decay is quadratic, but not cubic. For the left hexagon, cubic convergence occurs in the long run because the sequence $\tilde{F}^k_\omega(T)$ stays close to this intersection only for a few values of $k$, illustrating Theorem 3.

In the case $\Lambda = \mathrm{diag}(-1, 0, 1)$, the bottom edge consists of fixed points. This gives rise to a special asymptotic behavior ([8]): the (fixed) point labeled by $(0, 0, 0)$ is actually the limit of a collection of sequences $(\tilde{F}^k_\omega(T))$ for which the convergence is strictly quadratic.

# 4 Tubular coordinates

We collect a few basic facts about shifted $QR$ steps.

**Proposition 4.1** *If $s$ is not an eigenvalue of $\Lambda$, the map $F_s : \mathcal{T}_\Lambda \to \mathcal{T}_\Lambda$ is a diffeomorphism. The image of $F_{\lambda_i} : \mathcal{D}^i_\Lambda \to \mathcal{T}_\Lambda$ is $\mathcal{D}^i_{\Lambda,0}$. The restriction $F_{\lambda_i}|_{\mathcal{D}^i_{\Lambda,0}} : \mathcal{D}^i_{\Lambda,0} \to \mathcal{D}^i_{\Lambda,0}$ is a diffeomorphism.*

**Proof:** If $s$ is not an eigenvalue, compute $F_s^{-1}(T)$ by factoring $T - sI$ as $RQ$, $R$ upper triangular with the first $n-1$ diagonal entries positive and $Q \in SO(n)$: we claim that $F_s(T_0) = T$ for $T_0 = QR + sI$, proving that $F_s$ is a diffeomorphism. Indeed, $QR = T_0 - sI$ is a $Q_\star R_\star$ factorization and thus $F_s(T_0) = Q^* T_0 Q = T$.

From the last sentence of Section 2, the image of $F_{\lambda_i}$ is contained in $\mathcal{D}^i_{\Lambda,0} \subset \mathcal{D}^i_\Lambda$. The fact that the restriction of $F_{\lambda_i}$ to $\mathcal{D}^i_{\Lambda,0}$ is a diffeomorphism is proved as in the previous paragraph. ■

Commutativity of steps is well known and related to the complete integrability of the interpolating Toda flows ([5], [10], [12], [13]). For the reader's convenience we provide a proof.

**Proposition 4.2** *Steps commute: $F_{s_0} \circ F_{s_1} = F_{s_1} \circ F_{s_0}$ in the appropriate domains.*

The domain of $F_{s_0} \circ F_{s_1} = F_{s_1} \circ F_{s_0}$ is $\mathcal{T}_\Lambda$ if neither $s_0$ nor $s_1$ is an eigenvalue, $\mathcal{D}^i_\Lambda$ if $s_0 = \lambda_i$ and $s_1$ is not an eigenvalue (or vice-versa) and the empty set in the rather pointless case $s_0 = \lambda_i$, $s_1 = \lambda_j$, $i \neq j$.

**Proof:** We prove commutativity only when $s_0$ and $s_1$ are not eigenvalues; the other cases follow easily. Consider $Q_\star R_\star$ factorizations

$$T - s_0 I = Q_0 R_0, \quad T - s_1 I = Q_1 R_1,$$
$$(T - s_0 I)(T - s_1 I) = (T - s_1 I)(T - s_0 I) = Q_2 R_2.$$

For $F_{s_0}(T) - s_1 = Q_0^*(T - s_1)Q_0 = Q_3 R_3$, we have $F_{s_1}(F_{s_0}(T)) = Q_3^* F_{s_0}(T) Q_3 = Q_3^* Q_0^* T Q_0 Q_3$. Thus

$$Q_0^*(T - s_1)Q_0 R_0 = Q_0^*(T - s_1 I)(T - s_0 I) = Q_0^* Q_2 R_2 = Q_3 R_3 R_0$$

and therefore $Q_0^* Q_2 = Q_3$ and $F_{s_0}(F_{s_1}(T)) = Q_2^* T Q_2$. ■

Recall that a map $\Pi : X \to Y \subset X$ is a *projection* if $\Pi(X) = Y$ and $\Pi \circ \Pi = \Pi$. The map $F_{\lambda_i} : \mathcal{D}^i_\Lambda \to \mathcal{D}^i_{\Lambda,0}$ is not a projection but can be used to define one: the *canonical projection* $\Pi_i : \mathcal{D}^i_\Lambda \to \mathcal{D}^i_{\Lambda,0}$,

$$\Pi_i(T) = (F_{\lambda_i}|_{\mathcal{D}^i_{\Lambda,0}})^{-1}(F_{\lambda_i}(T)).$$

**Proposition 4.3** *The map $\Pi_i$ is indeed a smooth projection which commutes with steps. More precisely, $\Pi_i(F_s(T)) = F_s(\Pi_i(T))$ provided $s$ is not an eigenvalue of $\Lambda$ different from $\lambda_i$.*

**Proof:** The map $\Pi_i$ is clearly smooth and, for $T \in \mathcal{D}^i_{\Lambda,0}$, we have

$$\Pi_i(T) = (F_{\lambda_i}|_{\mathcal{D}^i_{\Lambda,0}})^{-1}(F_{\lambda_i}(T)) = T,$$

proving that $\Pi_i$ is a projection. Commutativity follows from Proposition 4.2. ■

For a diagonal matrix $\Lambda$ with simple spectrum and $\epsilon > 0$, the *deflation neighborhood* $\mathcal{D}_{\Lambda,\epsilon} \subset \mathcal{T}_\Lambda$ is the closed set of matrices $T \in \mathcal{T}_\Lambda$ with $|b(T)| \le \epsilon$. This notation is consistent with $\mathcal{D}_{\Lambda,0}$ for the deflation set. As we shall see in Propositions 4.4 and 5.1, for sufficiently small $\epsilon > 0$ the set $\mathcal{D}_{\Lambda,\epsilon}$ has connected components $\mathcal{D}^i_{\Lambda,\epsilon} \subset \mathcal{D}^i_\Lambda$, $\mathcal{D}^i_{\Lambda,\epsilon} \supset \mathcal{D}^i_{\Lambda,0}$, which are invariant under steps $F_s$ for shifts $s$ near $\lambda_i$, i.e., $F_s(\mathcal{D}^i_{\Lambda,\epsilon}) \subset \mathcal{D}^i_{\Lambda,\epsilon}$. The sets $\mathcal{D}^i_{\Lambda,\epsilon}$ are therefore also invariant under $F_\sigma$.

Denote the distance between a matrix $T$ and a compact set of matrices $\mathcal{N}$ by $\mathrm{dist}(T,\mathcal{N}) = \min_{S\in\mathcal{N}} \|T - S\|$. Let $\gamma = \min_{i\neq j} |\lambda_i - \lambda_j|$ be the *spectral gap* of $\Lambda$ and $B = e_n e^*_{n-1} + e_{n-1} e^*_n$.

Recall that if $\mathcal{N}$ is a submanifold of codimension $k$ of $\mathcal{M}$ then a *closed tubular neighborhood* of $\mathcal{N}$ consists of a closed neighborhood $\mathcal{N}_\epsilon$ of $\mathcal{N}$ and a diffeomorphism $\zeta : \mathcal{N}_\epsilon \to \mathcal{N} \times \mathbb{B}^k_\epsilon$ with $\zeta(x) = (x, 0)$ for $x \in \mathcal{N}$ (here $\mathbb{B}^k_\epsilon \subset \mathbb{R}^k$ is the closed ball of radius $\epsilon$ around the origin). Given $x \in \mathcal{N}$, the preimage $\zeta^{-1}(\{x\} \times \mathbb{B}^k_\epsilon)$ is a manifold with boundary of dimension $k$, the *fiber* through $x$. We now construct tubular neighborhoods of the deflation sets $\mathcal{D}^i_{\Lambda,0}$; here the codimension is $k = 1$.

**Proposition 4.4** *Each $\mathcal{D}^i_{\Lambda,0} \subset \mathcal{T}_\Lambda$ is a compact submanifold of codimension 1 diffeomorphic to $\mathcal{T}_{\Lambda_i}$, where $\Lambda_i = \mathrm{diag}(\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots \lambda_n)$. There exists $\epsilon_{\mathrm{tub}} > 0$ such that for $\epsilon \in (0, \epsilon_{\mathrm{tub}})$:*

*(a) the connected components $\mathcal{D}^i_{\Lambda,\epsilon}$ of $\mathcal{D}_{\Lambda,\epsilon}$ consist of matrices $T \in \mathcal{D}_{\Lambda,\epsilon}$ for which $|(T)_{n,n} - \lambda_i| < \sqrt{2}\,\epsilon$;*

*(b) the map $\zeta : \mathcal{D}^i_{\Lambda,\epsilon} \to \mathcal{D}^i_{\Lambda,0} \times [-\epsilon, \epsilon]$ given by $\zeta(T) = (\Pi_i(T), b(T))$ is a closed tubular neighborhood of $\mathcal{D}_{\Lambda,0}$;*

*(c) there is a constant $C_b > 0$ such that for all $T \in \mathcal{D}^i_{\Lambda,\epsilon}$,*

$$|b(T)| \le \mathrm{dist}(T, \mathcal{D}^i_{\Lambda,0}) \le \|T - \Pi_i(T)\| \le C_b |b(T)|.$$

**Proof:** We first show that the gradient of the restriction $b|_{\mathcal{T}_\Lambda}$ at a point $T_\mathcal{D} \in \mathcal{D}_{\Lambda,0}$ is not zero. Consider the characteristic polynomial along the line $T_\mathcal{D} + tB$: this is a smooth even function of $t$ and therefore $B$ is tangent to $\mathcal{T}_\Lambda$ at $T_\mathcal{D}$, the point on which $t = 0$. On the other hand, the directional derivative of $b$ along the same line equals 1. Thus $\mathcal{D}_{\Lambda,0} \subset \mathcal{T}_\Lambda$ is a submanifold of codimension 1. The diffeomorphism with $\mathcal{T}_{\Lambda_i}$ takes $T$ to $\hat{T}$, the leading $(n-1) \times (n-1)$ principal minor of $T$.

Assume $\epsilon < \gamma/(2\sqrt{2})$. Consider matrices $T \in \mathcal{D}_{\Lambda,\epsilon}$ and $S = T - b(T)B$, so that $(T)_{n,n}$ is an eigenvalue of $S$. By the Wielandt-Hoffman theorem, there exists an index $i$ for which $|(T)_{n,n} - \lambda_i| < \sqrt{2}\epsilon$, defining the sets $\mathcal{D}^i_{\Lambda,\epsilon}$ (at this point we do not yet know that $\mathcal{D}^i_{\Lambda,\epsilon}$ is connected).

For $T_\mathcal{D} \in \mathcal{D}^i_{\Lambda,0}$, the derivative $D\Pi_i(T_\mathcal{D})$ equals the identity on the subspace tangent to $\mathcal{D}^i_{\Lambda,0}$ and has a kernel of dimension 1. Thus, for sufficiently small $\epsilon_{\mathrm{tub}}$, item (b) holds. This also proves that each $\mathcal{D}^i_{\Lambda,\epsilon}$ is connected, completing the proof of item (a).

The first two inequalities in (c) are trivial. Now

$$\|T - \Pi_i(T)\| = \|\zeta^{-1}(\Pi_i(T), b(T)) - \zeta^{-1}(\Pi_i(T), 0)\| \le C_b |b(T)|,$$

where the derivative of $\zeta^{-1}(T_\mathcal{D}, \delta)$ with respect to the second coordinate is bounded by $C_b$ on the compact set $\mathcal{D}_{\Lambda,0} \times [-\epsilon_{\mathrm{tub}}, \epsilon_{\mathrm{tub}}]$. ■

The diffeomeorphism $\zeta$ defines *tubular coordinates* for $T \in \mathcal{D}^i_{\Lambda,\epsilon}$: the matrix $\Pi_i(T) \in \mathcal{D}^i_{\Lambda,0} \approx \mathcal{T}_{\Lambda_i}$ and $b(T)$. Under tubular coordinates, $QR$ steps with shift are given by a simple formula.

**Corollary 4.5** *Consider $\Lambda$, $i$ and $\epsilon \in (0, \epsilon_{\text{tub}})$. Then*

$$\zeta \circ F_s \circ \zeta^{-1} : \mathcal{D}^i_{\Lambda,0} \times [-\epsilon, \epsilon] \to \mathcal{D}^i_{\Lambda,0} \times [-\epsilon, \epsilon]$$
$$(T, b) \mapsto \left( F_s(T), \frac{(R_\star)_{n,n}}{(R_\star)_{n-1,n-1}}\, b \right)$$

*where $\zeta^{-1}(T, b) - sI = Q_\star R_\star$.*

**Proof:** This follows directly from Lemma 2.2 and Propositions 4.3 and 4.4. ■

# 5 Convergence to deflation

Sufficiently thin deflation neighborhoods $\mathcal{D}^i_{\Lambda,\epsilon}$ are invariant under $F_s$ for $s \approx \lambda_i$.

**Proposition 5.1** *Given $C > 0$, there exists $\epsilon_{\text{inv}} \in (0, \epsilon_{\text{tub}})$, such that for any $\epsilon \in (0, \epsilon_{\text{inv}})$ and $s \in [\lambda_i - C\,\epsilon, \lambda_i + C\,\epsilon]$ we have $F_s(\mathcal{D}^i_{\Lambda,\epsilon}) \subset \text{int}(\mathcal{D}^i_{\Lambda,\epsilon/2})$.*

*For a simple shift strategy $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$, there exists $\epsilon_{\text{inv}} > 0$ such that if $\epsilon \in (0, \epsilon_{\text{inv}})$ then $F_\sigma(\mathcal{D}^i_{\Lambda,\epsilon}) \subset \text{int}(\mathcal{D}^i_{\Lambda,\epsilon/2})$.*

In particular, $F_s$ is well defined in $\mathcal{D}^i_{\Lambda,\epsilon}$ for $\epsilon \in (0, \epsilon_{\text{inv}})$.

**Proof:** Recall that $F_s(\mathcal{D}^i_{\Lambda,0}) = \mathcal{D}^i_{\Lambda,0}$. From Lemma 2.3, the derivative of $b \circ \Phi_\star$ is zero at $\mathcal{D}^i_{\Lambda,0} \times \{\lambda_i\}$. Compactness of $\mathcal{D}^i_{\Lambda,0}$ thus implies that in a sufficiently small neighborhood of $\mathcal{D}^i_{\Lambda,0} \times \{\lambda_i\}$ we have $|b(F_s(T))| \leq |b(T)|/3$.

Now consider a simple shift strategy $\sigma$: by condition (II) there exists $C_\sigma > 0$ such that $|\sigma(T) - \lambda_i| < C_\sigma b(T)$; apply the first statement with $C = C_\sigma$. ■

Thus, $F_\sigma$ squeezes neighborhoods $\mathcal{D}^i_{\Lambda,\epsilon}$ at least linearly. Condition (I) and smoothness imply a stronger version of (II). We do not want to assume, however, that $\mathcal{D}_{\Lambda,0} \cap \mathcal{S}_\sigma = \emptyset$: after all, this is not true even for Wilkinson's shift. We need a more careful statement.

**Lemma 5.2** *Consider a shift strategy $\sigma$ and $\epsilon_{\text{inv}}$ as in Proposition 5.1. For a compact set $\mathcal{K} \subset \mathcal{D}^i_{\Lambda,\epsilon_{\text{inv}}} \smallsetminus (\mathcal{D}^i_{\Lambda,0} \cap \mathcal{S}_\sigma)$, there exists $C_\mathcal{K}$ such that for all $T \in \mathcal{K}$ we have $|\sigma(T) - \lambda_i| \leq C_\mathcal{K} b(T)^2$.*

**Proof:** Let $\mathcal{K}_\mathcal{D} = \mathcal{K} \cap \mathcal{D}^i_{\Lambda,0}$; enlarge $\mathcal{K}_\mathcal{D}$ along $\mathcal{D}^i_{\Lambda,0}$ to obtain another compact set $\mathcal{K}_1 \subset \mathcal{D}^i_{\Lambda,0} \smallsetminus \mathcal{S}_\sigma$, $\mathcal{K}_\mathcal{D} \subset \text{int}_{\mathcal{D}^i_{\Lambda,0}}(\mathcal{K}_1)$. Fatten $\mathcal{K}_1$ along fibers to define $\tilde{\mathcal{K}}_1 = \zeta^{-1}(\mathcal{K}_1 \times [-\epsilon, \epsilon])$, $\epsilon \in (0, \epsilon_{\text{inv}})$, which, without loss, still avoids $\mathcal{S}_\sigma$. For each $T_\mathcal{D} \in \mathcal{K}_1$, consider the function $h_{T_\mathcal{D}}(b) = \sigma(\zeta^{-1}(T_\mathcal{D}, b))$, obtained by restricting $\sigma$ to a fiber of $\mathcal{D}^i_{\Lambda,\epsilon}$. Each $h_{T_\mathcal{D}}$ is smooth and even (from condition (I)) and therefore satisfies $|h_{T_\mathcal{D}}(b) - \lambda_i| \leq C_{T_\mathcal{D}} |b|^2$. By compactness, there exists $C_{\mathcal{K}_1}$ such that $|h_{T_\mathcal{D}}(b) - \lambda_i| \leq C_{\mathcal{K}_1} |b|^2$ for all $T_\mathcal{D} \in \mathcal{K}_1$. In other words, there exists $C_{\tilde{\mathcal{K}}_1}$ such that $|\sigma(T) - \lambda_i| \leq C_{\tilde{\mathcal{K}}_1} |b(T)|^2$ for all $T \in \tilde{\mathcal{K}}_1$. The estimate for $T \notin \tilde{\mathcal{K}}_1$ is trivial. ■

**Proof of Theorem 2:** Take $\epsilon = \epsilon_{\text{inv}}$ as in Proposition 5.1 so that $\mathcal{D}^i_{\Lambda,\epsilon}$ is invariant under $F_\sigma$.

Let $\varphi = b \circ \Phi_\star$. We compute the Taylor expansion of $\varphi(T, s)$ at $(T_\mathcal{D}, \lambda_i)$, $T_\mathcal{D} \in \mathcal{D}^i_{\Lambda,0}$: from Lemma 2.3, the gradient of $\varphi$ at $(T_\mathcal{D}, \lambda_i)$ is zero. Thus, up to a third order remainder,

$$\begin{aligned}\varphi(T, s) = \varphi(T_\mathcal{D}, \lambda_i) &+ \frac{1}{2}\varphi_{T,T}(T_\mathcal{D}, \lambda_i)(T - T_\mathcal{D}, T - T_\mathcal{D}) + \\ &+ \varphi_{T,s}(T_\mathcal{D}, \lambda_i)(T - T_\mathcal{D}, s - \lambda_i) + \frac{1}{2}\varphi_{s,s}(T_\mathcal{D}, \lambda_i)(s - \lambda_i, s - \lambda_i) + \\ &+ \text{Rem}_3(T - T_\mathcal{D}, s - \lambda_i).\end{aligned}$$

Now, $\varphi(T_{\mathcal{D}}, \lambda_i) = 0$ and, again from Lemma 2.3, $\varphi(T, \lambda_i) = 0$ for all $T \in \mathcal{T}_\Lambda$, hence $\varphi_{T,T}(T_{\mathcal{D}}, \lambda_i) = 0$. Let $C_\sigma$ be the constant in condition (II) of the definition of a simple shift strategy. By compactness, there exists $C_1 > 0$ such that for all $T_{\mathcal{D}} \in \mathcal{D}^i_{\Lambda,0}$, $T \in \mathcal{D}^i_{\Lambda,\epsilon}$ and $s \in [\lambda_i - C_\sigma\,\epsilon, \lambda_i + C_\sigma\,\epsilon]$, we have

$$|\varphi(T, s)| \leq C_1 |s - \lambda_i| (\|T - T_{\mathcal{D}}\| + |s - \lambda_i|)$$

We now apply this estimate for $T_{\mathcal{D}} = \Pi_i(T)$, where $T \in \mathcal{D}^i_{\Lambda,\epsilon}$. By Proposition 4.4, since $\epsilon < \epsilon_{\text{tub}}$, $\|T - T_{\mathcal{D}}\| = \|T - \Pi_i(T)\| \leq C_b |b(T)|$ and therefore

$$|\varphi(T, s)| \leq C_1 |s - \lambda_i| (C_b |b(T)| + |s - \lambda_i|)$$

implying the quadratic estimate

$$|b(F_\sigma(T))| = |\varphi(T, \sigma(T))| \leq C_1 |\sigma(T) - \lambda_i| (C_b |b(T)| + |\sigma(T) - \lambda_i|) \leq C_q |b(T)|^2.$$

Using Lemma 5.2 instead of condition (II) yields the cubic estimate in (c). ■

As a corollary, we obtain the well-known fact that, near deflation, the rate of convergence of Rayleigh's shift is cubic. Similarly, the mixed Wilkinson-Rayleigh strategy has cubic convergence. The rate of convergence for Wilkinson's strategy is far subtler.

# 6 Deflationary strategies

On the way to prove Theorem 1, we construct a larger invariant set for $F_\sigma$. Let $\mathcal{U}_\Lambda \subset \mathcal{T}_\Lambda$ be the set of unreduced matrices; for $\epsilon > 0$, let $\mathcal{U}_{\Lambda,\epsilon} = \mathcal{U}_\Lambda \cup \text{int}(\mathcal{D}_{\Lambda,\epsilon})$. Notice that $\mathcal{U}_{\Lambda,\epsilon}$ is open, dense and path-connected.

**Lemma 6.1** *For a shift strategy $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$, $\epsilon_{\text{inv}}$ as in Proposition 5.1 and $\epsilon \in (0, \epsilon_{\text{inv}})$, the open set $\mathcal{U}_{\Lambda,\epsilon}$ is invariant under $F_\sigma$.*

**Proof:** If $T \in \mathcal{U}_\Lambda$ and $\sigma(T)$ is not in the spectrum then $F_\sigma(T)$ is (well defined and) unreduced. If $T \in \mathcal{U}_\Lambda$ and $\sigma(T) = \lambda_i$ then $F_\sigma(T) \in \mathcal{D}^i_{\Lambda,0} \subset \mathcal{U}_{\Lambda,\epsilon}$. Finally, if $T \in \text{int}(\mathcal{D}^i_{\Lambda,\epsilon})$ then, by Proposition 5.1, $F_\sigma(T) \in \text{int}(\mathcal{D}^i_{\Lambda,\epsilon/2}) \subset \mathcal{U}_{\Lambda,\epsilon}$. ■

Notice that we do not assume $\sigma$ or $F_\sigma$ to be continuous.

A simple shift strategy $\sigma$ is *deflationary* if for any $T \in \mathcal{U}_{\Lambda,\epsilon_{\text{inv}}}$ there exists $K \in \mathbb{N}$ such that $F^K_\sigma(T) \in \mathcal{D}_{\Lambda,\epsilon_{\text{inv}}}$.

Rayleigh's strategy is known not to be deflationary. The following well known estimate ([6] and [13], section 8-10) implies that Wilkinson's strategy is not only deflationary but uniformly so, in the sense that there exists $K$ with $F^K_\omega(\mathcal{U}_{\Lambda,\epsilon_{\text{inv}}}) \subset \mathcal{D}_{\Lambda,\epsilon_{\text{inv}}}$. As a corollary, the mixed Wilkinson-Rayleigh strategy is also uniformly deflationary provided $\epsilon > 0$ is sufficiently small.

**Fact 6.2** *For $T \in \mathcal{T}$ and $k \in \mathbb{N}$,*

$$|b(F^k_\omega(T))|^3 \leq \frac{|b(T)^2(T)_{n-1,n-2}|}{(\sqrt{2})^{k-1}}.$$

In [13], the result is shown for unreduced matrices; the case $T \in \mathcal{U}_{\Lambda,\epsilon_{\text{inv}}}$ follows by elementary limiting arguments. Notice that for $T \in \mathcal{T}_\Lambda$, the numerator $|b(T)^2(T)_{n-1,n-2}|$ is uniformly bounded.

We now prove Theorem 1: if the shift strategy $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$ is continuous then it is not deflationary.

**Proof of Theorem 1:** Fix $\epsilon = \epsilon_{\text{inv}}/2$. Let $\mathcal{B}^i \subset \mathcal{U}_{\Lambda,\epsilon}$ be the basins of attraction of each invariant neighborhood $\mathcal{D}^i_{\Lambda,\epsilon}$, i.e., $T \in \mathcal{B}^i$ if there exists $k \in \mathbb{N}$ such that $F^k_\sigma(T) \in \mathcal{D}^i_{\Lambda,\epsilon}$. The sets $\mathcal{B}^i$ are clearly disjoint with $\mathcal{D}^i_{\Lambda,\epsilon} \subset \mathcal{B}^i$. If $\sigma$ is continuous, they are also open subsets of $\mathcal{U}_{\Lambda,\epsilon}$ since $\mathcal{B}_i = \bigcup_k F^{-k}_\sigma(\text{int}(\mathcal{D}^i_{\Lambda,\epsilon}))$. If $\sigma$ is deflationary, $\bigcup_i \mathcal{B}_i = \mathcal{U}_{\Lambda,\epsilon}$. Thus, if $\sigma$ is both continuous and deflationary then $\mathcal{U}_{\Lambda,\epsilon}$ is not connected, a contradiction. ■

# 7 Dynamics of shifts for a.p. free matrices

From the previous section, cubic convergence may be lost when the orbit $F^k_\sigma(T)$ passes near the set $\mathcal{S}_\sigma \cap \mathcal{D}_{\Lambda,0}$. Our next task is to measure when this happens, by studying the dynamics associated to a shift strategy in a deflation neighborhood, i.e., the iterates of $F_\sigma : \mathcal{D}^i_{\Lambda,\epsilon} \to \mathcal{D}^i_{\Lambda,\epsilon}$, $\epsilon \in (0, \epsilon_{\text{inv}})$. Most of what we need can be read in the projection onto $\mathcal{D}^i_{\Lambda,0}$, where $F_\sigma$ coincides with $F_{\lambda_i}$.

A matrix $T \in \mathcal{T}$ with simple spectrum is *a.p. free* if no three eigenvalues are in arithmetic progression and a.p. otherwise. Different kinds of spectra lead to different dynamics: in this section we handle the a.p. free case, clearly a generic restriction. Let $\tilde{T}$ be the leading principal $(n-1)\times(n-1)$ minor of $T$. The following result is standard.

**Proposition 7.1** *Let $\Lambda \in \mathcal{T}$ be an $n \times n$ diagonal a.p. free matrix with spectrum $\lambda_1 < \cdots < \lambda_n$. For each $i$, consider $F_{\lambda_i} : \mathcal{D}^i_{\Lambda,0} \to \mathcal{D}^i_{\Lambda,0}$ as above. For any $T \in \mathcal{D}^i_{\Lambda,0}$, the sequence $(F^k_{\lambda_i}(T))$ converges to a diagonal matrix.*

**Proof:** The map $F_{\lambda_i}$ on $\mathcal{D}^i_{\Lambda,0}$ amounts to a $QR$ step with shift $\lambda_i$ on $\tilde{T}$, which has eigenvalues $\lambda_j$, $j \neq i$. The a.p. free hypothesis implies that the absolute values of the eigenvalues of $\tilde{T} - \lambda_i I$ are distinct. If $\tilde{T}$ is unreduced then, as is well known, the standard $QR$ iteration converges to a diagonal matrix, with diagonal entries in decreasing order of absolute value. More generally, if $\tilde{T}$ is reduced, apply the above result to each unreduced sub-block. ■

We shall use *height functions* for the $QR$ steps $F_s$, $s$ near $\lambda_i$, i.e., functions $H_i : \mathcal{D}^i_{\Lambda,\epsilon} \to \mathbb{R}$ with $H_i(F_s(T)) > H_i(T)$ provided $T$ is not diagonal. Such height functions and related scenarios have been considered in [1], [3], [10] and [14].

The matrix $W = \text{diag}(w_1, \ldots, w_n)$ is a *weight matrix* if $w_1 > \cdots > w_n$. Since $\Lambda$ is a.p. free, there exists $\epsilon_{\text{ap}} \in (0, \epsilon_{\text{inv}})$ such that if $s \in \mathcal{I}_i = [\lambda_i - \epsilon_{\text{ap}}, \lambda_i + \epsilon_{\text{ap}}]$ then the numbers $|\lambda_j - s|$ are distinct and their order does not depend on $s$.

**Proposition 7.2** *Let $\Lambda$ be an a.p. free diagonal matrix, $W$ a weight matrix and $\epsilon_{\text{ap}}$ as above. For $\delta_H > 0$, set $\eta_i(x) = \log((x - \lambda_i)^2 + \delta_H)$ and let $H_i : \mathcal{D}^i_{\Lambda,\epsilon_{\text{ap}}} \to \mathbb{R}$ be defined by $H_i(T) = \text{tr}(W\eta_i(T))$. There exists $\delta_H > 0$ such that*

$$\max_{T \in \partial\mathcal{D}^i_{\Lambda,\epsilon_{\text{ap}}}} H_i(T) < \min_{T \in \mathcal{D}^i_{\Lambda,0}} H_i(T)$$

*and, for any $s \in \mathcal{I}_i$, $H_i$ is a height function for $F_s : \mathcal{D}^i_{\Lambda,\epsilon_{\text{ap}}} \to \mathcal{D}^i_{\Lambda,\epsilon_{\text{ap}}}$.*

Here, $\eta_i(T) = X\,\text{diag}(\eta_i(\lambda_1), \ldots, \eta_i(\lambda_n))X^{-1}$ for $T = X\Lambda X^{-1}$ so that if $p$ is a polynomial and $\eta_i(\lambda_j) = p(\lambda_j)$ for $j = 1, \ldots, n$ then $\eta_i(T) = p(T)$. The only conditions on $\eta_i$ which will be used in the proof are that $|\lambda_j - \lambda_i| < |\lambda_k - \lambda_i|$ implies $\eta_i(\lambda_j) < \eta_i(\lambda_k)$ and that $\eta_i(\lambda_i)$ is very negative (for small $\delta_H$).

The proof requires some basic facts about *$f$-$Q_\star R_\star$ steps*; these facts will not be used elsewhere. For a real diagonal matrix $\Lambda$ with simple spectrum, let $\mathcal{O}_\Lambda$ be

the set of all real symmetric matrices similar to $\Lambda$; it is well known that $\mathcal{O}_\Lambda$ is a smooth compact manifold. The *f-$Q_\star R_\star$ step* applied to a matrix $S \in \mathcal{O}_\Lambda$ is the map $F_f : \mathcal{A}_{\Lambda,f} \to \mathcal{O}_\Lambda$ defined by $F_f(S) = Q_\star^* S Q_\star$, where $Q_\star$ is obtained from the factorization $f(S) = Q_\star R_\star$ and $S \in \mathcal{A}_{\Lambda,f}$ if and only if $f(S)$ is almost invertible. If $T \in \mathcal{T}_\Lambda \cap \mathcal{A}_{\Lambda,f}$ then $F_f(T) \in \mathcal{T}_\Lambda$ (use the same proof as in Lemma 2.2). The maps $F_s : \mathcal{T}_\Lambda \to \mathcal{T}_\Lambda$ defined above correspond to restrictions of $F_f$ for $f(x) = x - s$.

For a continuous function $h : \mathbb{R} \to \mathbb{R}$, if $S \in \mathcal{O}_\Lambda$ then the matrix function $h(S)$ belongs to $\mathcal{O}_M$, where $M = h(\Lambda)$. With the obvious abuse of notation, we have a diffeomorphism $h : \mathcal{O}_\Lambda \to \mathcal{O}_M$ provided $h$ is injective in the spectrum of $\Lambda$.

**Lemma 7.3** *For $h$ injective in the spectrum of $\Lambda$, consider the diffeomorphism $h : \mathcal{O}_\Lambda \to \mathcal{O}_M$, where $M = h(\Lambda)$. Let $f$ and $\tilde{f}$ be continuous functions defined in neighborhoods of the spectra of $\Lambda$ and $M$, respectively, satisfying $\tilde{f}(h(\lambda_j)) = f(\lambda_j)$ for each $j$ with $QR$ steps $F_f : \mathcal{O}_\Lambda \to \mathcal{O}_\Lambda$ and $F_{\tilde{f}} : \mathcal{O}_M \to \mathcal{O}_M$. Then $h \circ F_f = F_{\tilde{f}} \circ h$.*

**Proof:** The hypothesis implies that, for $T \in \mathcal{O}_\Lambda$, $f(T) = \tilde{f}(h(T)) = QR$ and hence $F_f(T) = Q^* T Q$ and $F_{\tilde{f}}(h(T)) = Q^* h(T) Q$. Thus $h(F_f(T)) = F_{\tilde{f}}(h(T))$. ■

Let $I_r$ be the $n \times n$ truncated identity matrix, i.e., $(I_r)_{i,i} = 1$ for $i \le r$, other entries being equal to zero.

**Lemma 7.4** *Let $M$ be a diagonal matrix with simple spectrum and $\tilde{f} : \mathbb{R} \to \mathbb{R}$ be a function for which $\mu_i < \mu_j$ implies $|\tilde{f}(\mu_i)| < |\tilde{f}(\mu_j)|$. Consider the $\tilde{f}$-$QR$ step $F_{\tilde{f}} : \mathcal{A}_{M,\tilde{f}} \to \mathcal{O}_M$. For any $S \in \mathcal{A}_{M,\tilde{f}}$ and $r = 1, \ldots, n-1$, $\operatorname{tr}(I_r F_{\tilde{f}}(S)) \ge \operatorname{tr}(I_r S)$. For $r = 1$, equality only holds if $(S)_{1,j} = 0$ for all $j > 1$.*

This argument follows closely the first proof in [3].

**Proof:** Let $V_r$ be the range of $I_r$ and $\mu_{r,j}(S)$ be the eigenvalues of the leading principal $r \times r$ minor of $S$, listed in nondecreasing order. We claim that $\mu_{r,j}(F_{\tilde{f}}(S)) \ge \mu_{r,j}(S)$, which immediately implies $\operatorname{tr}(I_r F_{\tilde{f}}(S)) \ge \operatorname{tr}(I_r S)$. Recall that $F_{\tilde{f}}(S) = Q_\star^* S Q_\star$ where $Q_\star R_\star = \tilde{f}(S)$. Let $U$ be an upper triangular matrix such that $Q_\star u = \tilde{f}(S) U u$ for $u \in V_r$. By min-max,

$$\mu_{r,j}(S) = \max_{\substack{A \subset V_r \\ \dim(A) = r+1-j}} \min_{u \in A \setminus \{0\}} \frac{\langle u, Su \rangle}{\langle u, u \rangle},$$

$$\mu_{r,j}(F_{\tilde{f}}(S)) = \max_A \min_u \frac{\langle u, F_{\tilde{f}}(S) u \rangle}{\langle u, u \rangle} = \max_A \min_u \frac{\langle \tilde{f}(S)Uu, S\tilde{f}(S)Uu \rangle}{\langle \tilde{f}(S)Uu, \tilde{f}(S)Uu \rangle}$$

$$= \max_{A' = UA} \min_{u' \in A' \setminus \{0\}} \frac{\langle \tilde{f}(S)u', S\tilde{f}(S)u' \rangle}{\langle \tilde{f}(S)u', \tilde{f}(S)u' \rangle}$$

Notice that since $U$ is upper triangular, the map taking $A \subset V_r$ to $A' = UA$ is a bijection among subspaces of $V_r$ of given dimension. Since $S$ and $\tilde{f}(S)$ are symmetric and commute,

$$\mu_{r,j}(F_{\tilde{f}}(S)) = \max_A \min_u \frac{\langle u, Sg(S)u \rangle}{\langle u, g(S)u \rangle},$$

where $g(x) = (\tilde{f}(x))^2$. The claim now follows from the inequality

$$\langle u, u \rangle \langle u, Sg(S)u \rangle - \langle u, Su \rangle \langle u, g(S)u \rangle \ge 0.$$

Diagonalize $S = Q^* M Q$ and $g(S) = Q^* g(\Lambda) Q$ and write $Qu = (x_1, \ldots, x_n)$ so that

$$2\left(\langle u, u \rangle \langle u, Sg(S)u \rangle - \langle u, Su \rangle \langle u, g(S)u \rangle\right) = \sum_{k,\ell} (\mu_k - \mu_\ell)(g(\mu_k) - g(\mu_\ell)) x_k^2 x_\ell^2 \ge 0.$$

Consider now equality for the case $r = 1$. Notice that, by hypothesis, if $k \neq \ell$ then $(\mu_k - \mu_\ell)(g(\mu_k) - g(\mu_\ell)) > 0$. In the max-min formula for $\mathrm{tr}(I_1 S) = \mu_{1,1}(S)$, it suffices to take $u = e_1$. Equality therefore holds only if $Qe_1$ is a canonical vector, which implies $(S)_{1,j} = 0$ for all $j > 1$. ■

**Proof of Proposition 7.2:** For all $s \in \mathcal{I}_i$ and any distinct eigenvalues $\lambda_j$ and $\lambda_k$, $|\lambda_j - \lambda_i| < |\lambda_k - \lambda_i|$ if and only if $\eta_i(\lambda_j) < \eta_i(\lambda_k)$. For $s \in \mathcal{I}_i$, $f(x) = x - s$, $h(x) = \eta_i(x)$ and $\mu_j = \eta_i(\lambda_j)$, define $\tilde{f} : \mathbb{R} \to \mathbb{R}$ as in Lemma 7.3. The function $\tilde{f}$ satisfies the hypothesis of Lemma 7.4: $\mu_j < \mu_k$ implies $|\tilde{f}(\mu_j)| < |\tilde{f}(\mu_k)|$. Thus, by Lemma 7.4, $\mathrm{tr}(WF_{\tilde{f}}(S)) \geq \mathrm{tr}(WS)$ for all $S \in \mathcal{O}_\Sigma$. For $T \in \mathcal{D}^i_{\Lambda,\epsilon_{\mathrm{ap}}}$, take $S = h(T)$: by Lemma 7.3, $F_{\tilde{f}}(h(T)) = h(F_f(T))$ and therefore $\mathrm{tr}(Wh(F_f(T))) \geq \mathrm{tr}(Wh(T))$. Again by Lemma 7.4, equality happens only if $T$ is diagonal. Thus, $H_i$ is a height function. Finally, choosing $\delta_H$ sufficiently small guarantees that $H_i$ is large in $\mathcal{D}^i_{\Lambda,0}$ and small in $\partial\mathcal{D}^i_{\Lambda,\epsilon_{\mathrm{ap}}}$, completing the proof. ■

Thus, simple shift strategies admit height functions near the deflation set. Our reason for constructing a height function is to control the time the sequence $(F^k_\sigma(T))$ stays in a compact set.

Assuming $\Lambda$ to be a.p. free, for a shift strategy $\sigma : \mathcal{T}_\Lambda \to \mathbb{R}$ set $\epsilon_\sigma = \epsilon_{\mathrm{ap}}/(1 + C_\sigma)$ (where $C_\sigma$ is the constant in the definition of a simple shift strategy). Notice that $T \in \mathcal{D}^i_{\Lambda,\epsilon_\sigma}$ implies $\sigma(T) \in \mathcal{I}_i = [\lambda_i - \epsilon_{\mathrm{ap}}, \lambda_i + \epsilon_{\mathrm{ap}}]$.

**Corollary 7.5** *Let $\Lambda$ be a real diagonal $n \times n$ a.p. free matrix, $\sigma$ a simple shift strategy and $\mathcal{D}^i_{\Lambda,\epsilon_\sigma}$ as above. Let $\mathcal{K} \subset \mathcal{D}^i_{\Lambda,\epsilon_\sigma}$ be a compact set with no diagonal matrices: there exists $K \in \mathbb{N}$ such that for all $T \in \mathcal{D}^i_{\Lambda,\epsilon_\sigma}$ there are at most $K$ points of the form $F^k_\sigma(T)$ in $\mathcal{K}$.*

The plan is to take $\mathcal{K}$ containing $\mathcal{S}_\sigma \cap \mathcal{D}^i_{\Lambda,\epsilon_\sigma}$: the hypothesis in Theorem 3 that diagonal matrices do not belong to the singular support $\mathcal{S}_\sigma$ is then natural.
**Proof:** Let $m_-$ be the minimum jump in $\mathcal{K}$ and $m_+$ the size of the image of $H_i$:

$$m_- = \inf_{T\in\mathcal{K},\, s\in\mathcal{I}_i} H_i(F_s(T)) - H_i(T), \quad m_+ = \sup_{T\in\mathcal{D}^i_{\Lambda,\epsilon_\sigma}} H_i(T) - \inf_{T\in\mathcal{D}^i_{\Lambda,\epsilon_\sigma}} H_i(T).$$

By Proposition 7.2 and the compactness of $\mathcal{K} \times \mathcal{I}_i$, $s > 0$: take $K$ such that $Km_- > m_+$. For a given $T$, let $X = \{k \in \mathbb{N} \mid F^k_\sigma(T) \in \mathcal{K}\}$: we have

$$m_+ \geq \sum_{k\in X} H_i(F^{k+1}_\sigma(T)) - H_i(F^k_\sigma(T)) \geq |X| m_-$$

and therefore $|X| < K$. ■

**Proof of Theorem 3:** Let $\mathcal{K}_1, \mathcal{K}_2 \subset \mathcal{D}^i_{\Lambda,\epsilon_\sigma}$ be compact sets with $\mathcal{K}_1 \cup \mathcal{K}_2 = \mathcal{D}^i_{\Lambda,\epsilon_\sigma}$, $\mathcal{S}_\sigma \cap \mathcal{D}^i_{\Lambda,0}$ disjoint from $\mathcal{K}_1$ and with no diagonal matrices in $\mathcal{K}_2$. By Theorem 2, there exists $C_{\mathcal{K}_1} > 0$ such that $|b(F_\sigma(T))| \leq C_{\mathcal{K}_1} |b(T)|^3$ for all $T \in \mathcal{K}_1$. By Corollary 7.5, there exists $K_2 \in \mathbb{N}$ such that, given $T \in \mathcal{D}^i_{\Lambda,\epsilon_\sigma}$, at most $K_2$ points of the form $F^k_\sigma(T)$ belong to $\mathcal{K}_2$. In particular, there are at most $K_2$ values of $k$ for which the estimate $|b(F^{k+1}_\sigma(T))| \leq C_{\mathcal{K}_1} |b(F^k_\sigma(T))|^3$ does not hold. ■

# 8 Convergence properties of a.p. spectra

The aim of this section is to prove Theorem 4. An a.p. matrix $T \in \mathcal{T}$ with simple spectrum is *strong a.p.* if three consecutive eigenvalues are in arithmetic progression and *weak a.p.* otherwise.

In the a.p. free case discussed in the previous sections, for an initial condition $T \in \mathcal{D}^i_{\Lambda,\epsilon}$, the sequence $F^k_\sigma(T)$ converges to a diagonal matrix; this follows from the fact that $\sigma(T) \approx \lambda_i$ for $T \in \mathcal{D}^i_{\Lambda,\epsilon}$. For weak a.p. spectra, convergence to a diagonal matrix may not occur.

Assume $\Lambda$ to be weak a.p. Let $b_2(T) = T_{n-1,n-2}$ be the second-last subdiagonal entry; for consistency, write $b_1(T) = b(T)$. For any $i$, there exists a unique index $c(i)$ such that $\lambda_{c(i)}$ is the eigenvalue closest to $\lambda_i$. As we shall see, if $T \in \mathcal{D}^i_{\Lambda,\epsilon}$ then

$$\lim_{k\to\infty} b_1(F^k_\sigma(T)) = \lim_{k\to\infty} b_2(F^k_\sigma(T)) = 0, \quad \lim_{k\to\infty} (F^k_\sigma(T))_{n,n} = \lambda_i;$$

furthermore, if $T$ is unreduced then

$$\lim_{k\to\infty} (F^k_\sigma(T))_{n-1,n-1} = \lambda_{c(i)}.$$

We begin with a technical lemma concerning the dynamics of steps $F_s$. Item (b) is a variation of the power method argument used to study the convergence of lower entries under $QR$ steps.

**Lemma 8.1** *Let $M = \mathrm{diag}(\mu_1, \ldots, \mu_m)$ be a real diagonal matrix with simple spectrum and $\mathcal{T}_M \subset \mathcal{T}$ be the manifold of real $m \times m$ tridiagonal matrices similar to $M$. Let $I \subset \mathbb{R}$ be a compact interval. Assume that there exists $j$, $1 \le j \le m$, such that*

$$\mu_j \notin I, \quad \max_{s\in I} |\mu_j - s| < \min_{k\neq j, s\in I} |\mu_k - s|.$$

*Let $\mathcal{D}^j_{M,\epsilon} \subset \mathcal{T}_M$ be the $j$-th deflation neighborhood.*

*(a) There exist $\epsilon > 0$ and $C \in (0,1)$ such that for all $\epsilon' \in (0,\epsilon)$ and $s \in I$ we have $F_s(\mathcal{D}^j_{M,\epsilon'}) \subset \mathcal{D}^j_{M,C\epsilon'}$.*

*(b) Consider $T_0 \in \mathcal{T}_M$ unreduced, a sequence $(s_k)$ of elements of $I$ and $\epsilon > 0$. Define $T_{k+1} = F_{s_k}(T_k)$. Then there exists $k$ such that $T_k \in \mathcal{D}^j_{M,\epsilon}$.*

This will be used to study $b_2(T)$ for $T \in \mathcal{D}^i_{\Lambda,\epsilon}$, setting $I = [\lambda_i - \epsilon, \lambda_i + \epsilon]$, $j = c(i)$, $M = \Lambda_i = \mathrm{diag}(\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots, \lambda_n)$, with the natural identification between $\mathcal{T}_M$ and $\mathcal{D}^i_{\Lambda,0}$.

**Proof:** Let $\tilde{C} \in (0,1)$ be such that

$$\max_{s\in I} |\mu_j - s| < \tilde{C} \min_{k\neq j, s\in I} |\mu_k - s|.$$

Write

$$r(s,T) = \frac{(R_\star)_{m,m}}{(R_\star)_{m-1,m-1}}, \quad T - sI = Q_\star R_\star.$$

Recall from Lemma 2.2 and Corollary 4.5 that $b(F_s(T)) = r(s,T)\, b(T)$. We claim that for all $T \in \mathcal{D}^j_{M,0}$ and $s \in I$, $|r(s,T)| \le \tilde{C}$. Since $T \in \mathcal{D}^j_{M,0}$, $|(R_\star)_{m,m}| = |\mu_j - s|$. Let $R_-$ be the leading principal minor of $R_\star$ of order $m-1$: its singular values are $|\mu_k - s|$, $k \neq s$. In particular, all singular values are larger that $|(R_\star)_{m,m}|/\tilde{C}$. Thus

$$|(R_\star)_{m-1,m-1}| = \|e^*_{m-1} R_-\| \ge \frac{|(R_\star)_{m,m}|}{\tilde{C}} \|e_{m-1}\| = \frac{|(R_\star)_{m,m}|}{\tilde{C}},$$

proving our claim. Take $C = (1 + \tilde{C})/2$: by continuity, for sufficiently small $\epsilon > 0$, we have $|r(s,T)| < C$ for all $T \in \mathcal{D}^j_{M,\epsilon}$, $s \in I$. Thus, for $T \in \mathcal{D}^j_{M,\epsilon}$ and $s \in I$, $|b(F_s(T))| \le C\, |b(T)|$; item (a) follows.

For item (b), write $T_{k+1} = Q_k^* T_k Q_k$ where $T_k - s_k I = Q_k R_k$ is a $Q_\star R_\star$ decomposition. Notice that, by hypothesis, $I$ is disjoint from the spectrum so that

$T_0 - s_0 I$ is invertible. We have $(T_0 - s_0 I)^{-1} = R^{-1} Q_0^\star$ so the rows of $Q_0^\star$ are obtained from those of $(T_0 - s_0 I)^{-1}$ by Gram-Schmidt from bottom to top. In particular, $Q_0 e_m = c_0 (T_0 - s_0 I)^{-1} e_m$, $c_0 > 0$. More generally, we claim that

$$P_k e_m = c(T_0 - s_{k-1} I)^{-1} \cdots (T_0 - s_1 I)^{-1} (T_0 - s_0 I)^{-1} e_m,$$
$$c > 0, \quad P_k = Q_0 Q_1 \cdots Q_{k-1} \in SO(m).$$

Indeed, by induction and using that $T_1 = Q_0^* T_0 Q_0$,

$$\begin{aligned} P_k e_m &= c' Q_0 (T_1 - s_{k-1} I)^{-1} \cdots (T_1 - s_1 I)^{-1} e_m \\ &= c' (T_0 - s_{k-1} I)^{-1} \cdots (T_0 - s_1 I)^{-1} Q_0 e_m \\ &= c (T_0 - s_{k-1} I)^{-1} \cdots (T_0 - s_1 I)^{-1} (T_0 - s_0 I)^{-1} e_m. \end{aligned}$$

For $\alpha = 1, \ldots, m$, let $v_\alpha$ be the unit eigenvector associated to $\mu_\alpha$. We claim that

$$\lim_{k \to \infty} P_k e_m = \pm v_j.$$

Indeed, write $e_m = \sum_{\alpha=1}^m a_\alpha v_\alpha$, where $a_\alpha = \langle v_\alpha, e_m \rangle$ is the last coordinate of $v_\alpha$. It is well known that the last coordinates of the eigenvectors $v_\alpha$ of the unreduced matrix $T$ are nonzero: in particular, $a_j \neq 0$; assume without loss $a_j > 0$. We have

$$\begin{aligned} P_k e_m &= c(T_0 - s_{k-1} I)^{-1} \cdots (T_0 - s_1 I)^{-1} (T_0 - s_0 I)^{-1} e_m \\ &= c \sum_{\alpha=1}^m \frac{a_\alpha}{(\mu_\alpha - s_{k-1}) \cdots (\mu_\alpha - s_0)} v_\alpha = c_k \left( v_j + \sum_{\alpha \neq j} b_{k,\alpha} v_\alpha \right), \\ & c_k > 0, \quad b_{k,\alpha} = \frac{a_\alpha}{a_j} \frac{\mu_j - s_{k-1}}{\mu_\alpha - s_{k-1}} \cdots \frac{\mu_j - s_0}{\mu_\alpha - s_0}. \end{aligned}$$

Since $|\mu_j - s_{k-1}| / |\mu_\alpha - s_{k-1}| < \tilde{C}$ we have $|b_{k,\alpha}| \leq (\tilde{C})^k |a_\alpha / a_j|$ and therefore $\lim_{k \to \infty} b_{k,\alpha} = 0$, proving the claim. We have

$$\lim_{k \to \infty} b(T_k) = \lim_{k \to \infty} (T_k)_{m,m-1} = \lim_{k \to \infty} e_{n-1}^* T_k e_m = \lim_{k \to \infty} (P_k e_{m-1})^* T_0 (P_k e_m) =$$
$$= \lim_{k \to \infty} (P_k e_{m-1})^* \mu_j (P_k e_m) + \lim_{k \to \infty} (P_k e_{m-1})^* (T_0 - \mu_j I)(P_k e_m).$$

The first limit in the last expression is zero because $P_k e_{m-1} \perp P_k e_m$; the second is zero because $P_k e_{m-1}$ is bounded and

$$\lim_{k \to \infty} (T_0 - \mu_j I)(P_k e_m) = (T_0 - \mu_j I) \lim_{k \to \infty} (P_k e_m) = (T_0 - \mu_j I) v_j = 0.$$

■

Consider the *double deflation set* $\mathcal{C}_{\Lambda,0} \subset \mathcal{D}_{\Lambda,0} \subset \mathcal{T}_\Lambda$:

$$\mathcal{C}_{\Lambda,0} = \{ T \in \mathcal{T}_\Lambda \mid b_1(T) = b_2(T) = 0 \}.$$

For Wilkinson's strategy $\omega$, it turns out that the set $\mathcal{C}_{\Lambda,0}$ is disjoint from the singular support $\mathcal{S}_\omega$. More generally, if a shift strategy $\sigma$ satisfies $\mathcal{C}_{\Lambda,0} \cap \mathcal{S}_\sigma = \emptyset$ then cubic convergence of $F_\sigma$ holds even for weak a.p. spectra: this is Theorem 4, which we prove below.

In [8], we show examples of unreduced tridiagonal $3 \times 3$ matrices with spectrum $-1, 0, 1$ for which Wilkinson's shift $F_\omega$ converges quadratically to a reduced but not diagonal matrix in the singular support $\mathcal{S}_\omega$. Similarly, we conjecture that for strong a.p. diagonal $n \times n$ matrices $\Lambda$ there exists a set $\mathcal{X} \subset \mathcal{T}_\Lambda$ of Hausdorff codimension 1 of unreduced matrices $T$ for which $F_\omega^k(T)$ converges quadratically to a matrix in $\mathcal{S}_\omega \cap \mathcal{D}_{\Lambda,0}$ with $T_{n-1,n-2} \neq 0$.

With the natural identification between $\mathcal{D}^i_{\Lambda,0}$ and $\mathcal{T}_{\Lambda_i}$, we may consider $\mathcal{D}^j_{\Lambda_i,\epsilon_2}$ to be a subset of $\mathcal{D}^i_{\Lambda,0}$. Let

$$\mathcal{C}^{j,i}_{\Lambda,\epsilon_2,\epsilon_1} = \mathcal{D}^i_{\Lambda,\epsilon_1} \cap \Pi_i^{-1}(\mathcal{D}^j_{\Lambda_i,\epsilon_2}).$$

For small $\epsilon_1, \epsilon_2 > 0$, $T \in \mathcal{C}^{j,i}_{\Lambda,\epsilon_2,\epsilon_1}$ implies

$$T_{n-1,n-1} \approx \lambda_j, \quad T_{n,n} \approx \lambda_i, \quad b_1(T) \leq \epsilon_1, \quad b_2(T) \approx 0.$$

These compact sets turn out to be manifolds with corners but we shall neither prove nor use this fact. Lemma 8.1 can be rephrased in terms of the sets $\mathcal{C}^{j,i}_{\Lambda,\epsilon_2,\epsilon_1}$.

**Corollary 8.2** *Let $\Lambda$ to be weak a.p. spectrum and $\sigma$ be a simple shift strategy. There exists $\epsilon > 0$ such that, for all $i$ and for all $\epsilon_1 \in (0, \epsilon)$:*

*(a) there exists $C \in (0,1)$ such that, for all sufficiently small $\epsilon_2 > 0$ we have $F_\sigma(\mathcal{C}^{c(i),i}_{\Lambda,\epsilon_2,\epsilon_1}) \subset \mathcal{C}^{c(i),i}_{\Lambda,C\epsilon_2,\epsilon_1}$;*

*(b) for all unreduced $T \in \mathcal{D}^i_{\Lambda,\epsilon}$ and for all $\epsilon_1, \epsilon_2 > 0$ there exists $k$ such that $F^k_\sigma(T) \in \mathcal{C}^{c(i),i}_{\Lambda,\epsilon_2,\epsilon_1}$.*

**Proof:** Combine Lemma 8.1 with $\Pi_i \circ F_s = F_s \circ \Pi_i$ (Proposition 4.3). ■

**Proof of Theorem 4:** From the hypothesis that $\mathcal{C}_{\Lambda,0}$ and $\mathcal{S}_\sigma$ are disjoint it follows that, for sufficiently small $\epsilon_1, \epsilon_2 > 0$, the shift strategy $\sigma$ is smooth in $\mathcal{C}^{c(i),i}_{\Lambda,\epsilon_2,\epsilon_1}$. As in Lemma 5.2, from a Taylor expansion around $T_0 \in \mathcal{D}^i_{\Lambda,0}$, there exists $C_2$ such that $|\sigma(T)| \leq C_2 |b_1(T)|^2$ for all $T \in \mathcal{C}^{c(i),i}_{\Lambda,\epsilon_2,\epsilon_1}$. As in the proof of Theorem 2, there exists $C_3$ such that $|b_1(F_\sigma(T))| \leq C_3 |b_1(T)|^3$ for all $T \in \mathcal{C}^{c(i),i}_{\Lambda,\epsilon_2,\epsilon_1}$. From item (a) of Corollary 8.2, $\mathcal{C}^{c(i),i}_{\Lambda,\epsilon_2,\epsilon_1}$ is invariant under $F_\sigma$; from item (b), for all unreduced $T \in \mathcal{D}^i_{\Lambda,\epsilon}$ (where $\epsilon$ is sufficiently small) there exists $K$ such that, for all $k > K$, $F^k_\sigma(T) \in \mathcal{C}^{c(i),i}_{\Lambda,\epsilon_2,\epsilon_1}$, completing the proof. ■

# 9 Two counterexamples

In this section we present two examples which show that natural strengthenings of Theorems 3 and 4 do not hold for Wilkinson's strategy $\omega$.

We use the notation of Section 3. In Figure 2, where $\Lambda = \mathrm{diag}(1,2,4)$, we indicate a sequence $\tilde{F}^k_\omega(T)$ which enters the deflation neighborhood $\mathcal{D}^i_{\Lambda,\epsilon}$ near one diagonal matrix but travels within the neighborhood towards another diagonal matrix. Theorem 2 guarantees the cubic decay of the $(3,2)$ entry whenever $\tilde{F}^k_\omega(T)$ stays away from the singular support $\mathcal{S}_\omega$. Consistently with Theorem 3, this happens for practically all values of $k$. Notice however that no uniform bound exists on the number of iterations needed to reach (a neighborhood of) $\mathcal{S}_\omega$. As proved in [8], in this instance cubic decay does not hold. More precisely, it is *not* true that given an a.p. free matrix $\Lambda$ there exist $C > 0$ and $K$ such that $|b(F^{k+1}_\omega(T))| \leq C |b(F^k_\omega(T))|^3$ for all $k > K$.

Consider now the weak a.p. spectrum $\Lambda = \mathrm{diag}(-1, 0, 0.3, 1)$ and

$$T_0 = \begin{pmatrix} 0.3 & 0 \\ 0 & S_0 \end{pmatrix} \in \mathcal{T}_\Lambda$$

where $S_0 \in \mathcal{T}_{\Lambda_3}$, $\Lambda_3 = \mathrm{diag}(-1,0,1)$, is an example of unreduced matrix obtained in [8] for which convergence is strictly quadratic, i.e.,

$$C_- |b(F^k_\omega(S_0))|^2 < |b(F^{k+1}_\omega(S_0))| < C_+ |b(F^k_\omega(S_0))|^2,$$

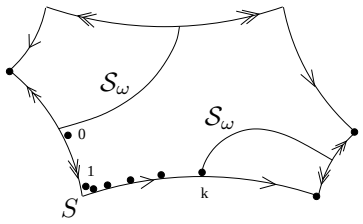


Figure 2: We may have $F_\omega^k(T) \in \mathcal{S}_\omega$ for large values of $k$.

for all $k$, where $0 < C_- < C_+$. Trivially, the analogous estimate holds for $b(F_\omega^k(T_0))$. By sheer continuity, given $K$, there exists $\epsilon > 0$ such that if $T \in \mathcal{T}_\Lambda$ satisfies $\|T - T_0\| < \epsilon$ then

$$C_-|b(F_\omega^k(T))|^2 < |b(F_\omega^{k+1}(T))| < C_+|b(F_\omega^k(T))|^2$$

still holds for all $k < K$. Thus, the uniform estimate in Theorem 3 fails for weak a.p. spectra, even for unreduced matrices.

# References


[1] Bloch, A. M., Brockett, R. W. and Ratiu, T., Completely integrable gradient flows, Comm. Math. Phys., 147, 57-74, 1992.

[2] Deift, P., Nanda, T., Tomei, C., Differential equations for the symmetric eigenvalue problem, SIAM J. Num. Anal. 20, 1-22, 1983.

[3] Deift, P., Rivera, S., Tomei, C., Watkins, D., A monotonicity property for Toda-type flows, SIAM J. Matrix Anal. Appl., 12, 463-468, 1991.

[4] Demmel, J. W., Applied Numerical Linear Algebra, SIAM, Philadelphia, 1997.

[5] Flaschka, H., The Toda lattice, I, Phys. Rev. B 9, 1924-1925, 1974.

[6] Hoffmann, W. and Parlett, B., A new proof of global convergence for the tridiagonal QL algorithm, SIAM J. Num. Anal. 15 (1978), 929-937.

[7] Leite, R. S., Saldanha, N.C. and Tomei, C., An atlas for tridiagonal isospectral manifolds, Lin. Alg. Appl. 429 (2008), 387-402.

[8] Leite, R. S., Saldanha, N.C. and Tomei, C., The Asymptotics of Wilkinson's shift: Loss of Cubic Convergence, Foundations of Computational Mathematics, to appear. doi:10.1007/s10208-009-9047-3

[9] Leite, R. S., Saldanha, N.C. and Tomei, C., The Asymptotics of Wilkinson's shift iteration, arXiv:math/0412493v2.

[10] Leite, R. S. and Tomei, C., Parametrization by polytopes of intersections of orbits by conjugation, Lin. Alg. and Appl. 361, 223-246, 2003.

[11] Moerbeke, P. van, The spectrum of Jacobi matrices, Inventiones Math., 37, 45-81, 1976.

[12] Moser, J., Finitely many mass points on the line under the influence of an exponential potential, In: Dynamic systems theory and applications, (ed. J. Moser) 467-497, New York, 1975.

[13] Parlett, B. N., The Symmetric Eigenvalue Problem, Classics in Applied Mathematics 20, SIAM, 1997.

[14] Tomei, C., The Topology of Manifolds of Isospectral Tridiagonal Matrices, Duke Math. J., 51, 981-996, 1984.

[15] Wilkinson, J. H., The algebraic eigenvalue problem, Oxford University Press, 1965.

Ricardo S. Leite, Departamento de Matemática, UFES
Av. Fernando Ferrari, 514, Vitória, ES 29075-910, Brazil

Nicolau C. Saldanha and Carlos Tomei, Departamento de Matemática, PUC-Rio
R. Marquês de S. Vicente 225, Rio de Janeiro, RJ 22453-900, Brazil

rsleite@cce.ufes.br
saldanha@puc-rio.br; http://www.mat.puc-rio.br/∼nicolau/
tomei@mat.puc-rio.br